© Journal of the American Statistical Association
June 1971, Volume 66, Number 334
Applications Section

Accident Rate Potential: An Application of Multiple Regression Analysis of a Poisson Process

DONALD C. WEBER*

Various accident frequency models have appeared in the literature which predict the distribution of future accidents based on the number of past accidents. This article presents a method for deriving such distributions using several predictive criteria. It is assumed that an individual's accident experience is a Poisson process with the parameter a linear function of criterion variables. An iterative weighted least-squares procedure is used to solve the system of maximum likelihood equations required for estimating this parameter and a large sample test procedure is illustrated. The tenability of the model is viewed in the light of actual data.

1. INTRODUCTION

In 1920, Greenwood and Yule [6] proposed an accident frequency model which assumes that the number of accidents experienced by an individual is a Poisson process, i.e., the probability of his experiencing n accidents during a time interval of length t, $p(n, t)$, is given by

$$p(n, t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \qquad \begin{matrix} n = 0, 1, 2, \cdots, \\ \lambda > 0, \quad t > 0, \end{matrix} \tag{1.1}$$

and assumes further that λ is a value of a random variable having a gamma distribution with density function

$$g(\lambda) = \frac{(r/m)^r}{\Gamma(r)} \lambda^{r-1} e^{-(r/m)\lambda}, \qquad \lambda > 0, \quad m > 0, \quad r > 0.$$

The resulting unconditional distribution for n accidents in time t is negative binomial

$$q(n, t) = \binom{n + r - 1}{r - 1} \left(\frac{r}{r + mt}\right)^r \left(\frac{mt}{r + mt}\right)^n, \tag{1.2}$$
$$n = 0, 1, 2, \cdots, \quad r > 0, \quad m > 0, \quad t > 0,$$

with mean mt and variance $mt(1+mt/r)$.

Using this model, Arbous and Kerrich [1], Bates and Neyman [2], and Edwards and Gurland [4] derived various bivariate accident distribution models which differ in certain underlying assumptions. These bivariate models can be used to obtain distributions of future accidents conditioned upon the number of past accidents. Insurance companies and motor vehicle departments have long recognized, however, that factors other than accident history are related to future automobile accident experience, e.g., age, sex, geographic location, mileage driven and conviction history. Accepting the proposition that an individual driver's accident frequency over a short period of time follows (1.1), the purpose of this article is to demonstrate a method for estimating the parameter λ, the "accident rate potential" associated with the individual, as a function of k criteria.

2. THE DATA

The 1964 California Driver Record Study [3] contains data collected on about 148,000 motorists over a period of three years, 1961–63. These data include information on certain attributes of the individuals in the sample along with their driving record. In Section 5, data from the California study will be used to arrive at an estimating function for λ with respect to male drivers during 1963 based on certain personal characteristics and driving performance of the individual during the preceding two years. In Table 1 we see the fit of the negative binomial model (1.2) to the sample data using the method of moments.

Table 1. ACTUAL AND THEORETICAL 1963 ACCIDENT DISTRIBUTIONS FOR MALE DRIVERS[a]

Accidents	Actual	Theoretical
0	80,369	80,372
1	5,910	5,902
2	415	420
3+	32	32
	86,726	86,726

$\chi^2 = 0.07$, 1 d.f.

[a] Source: [3, Part 6].

To explore further the validity of the model's assumptions, the 148,000 individuals in the California sample were partitioned into 2,880 groups according to sex, marital status, age, area of residence, conviction history and accident history. A computer program was used to obtain the 1963 accident distributions for the 193 groups that contained 100 or more individuals and to fit a Poisson distribution (1.1) to each such group. In 167 or 86.5 percent of the 193 cases the hypothesis of a Poisson

* Donald C. Weber is assistant professor of mathematics, Miami University
This research was supported in part under a National Institutes of Health Training Grant and is based on a portion of the author's doctoral dissertation submitted to North Carolina State University. The author is indebted to the California Department of Motor Vehicles for unrestricted use of their 1964 Driver Record Study [3] data.

distribution was acceptable at the .05 level of significance. We may conclude from these results that the six criterion variables did a credible job of classifying the individuals into Poisson groups.

3. THE MODEL

Additional analysis of the California data reveals that a linear or near linear relationship exists between the mean accident frequency per unit time and several investigated criterion variables. For example, a plot of accident rate versus age class (in 5-year units) suggests a hyperbolic type relationship between the two variables. If we assign the reciprocal of the integers 1, 2, · · · , 12 to the twelve age classes by the transformation

$$x = 5/(\text{age} - 13)$$

and let y represent the accident rate over the three-year period, we obtain the estimation function

$$y = 0.1823 + 0.3183x$$

using weighted regression analysis. In Table 2 we find the comparison of estimated accident rates with the empirical rates.

Table 2. EMPIRICAL AND ESTIMATED ACCIDENT RATES ON TRANSFORMED AGES FOR MALES, 1961–63[a]

Age class	Empirical	Estimated
Under 21	.468	.459
21 – 25	.332	.341
26 – 30	.290	.288
31 – 40	.253	.253
41 – 60	.229	.226
Over 60	.204	.210

[a] *Source*: [3, Part 5].

As a second illustration, we may look at the relationship between future accident involvements and accident history. If we accept the tenet that the negative binomial model (1.2) is at least an approximation to actual automobile experience, we have reason to believe that future accident rates are linearly related to the incidence of past accident involvements on a theoretical basis (see [1]). To confirm this, a weighted regression analysis of 1963 accidents on the number of 1961–62 involvements for male California motorists produced the rate function

$$y = 0.07234 + 0.03818x.$$

The result of this analysis is given in Table 3. Similarly, linear relationships were found between accident rates and other variables such as traffic density and conviction count.

On the basis of these analyses let us hypothesize that the parameter λ of (1.1) is a linear function of k classification or criterion variables, i.e.,

$$\lambda = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{3.1}$$

Table 3. EMPIRICAL AND ESTIMATED 1963 ACCIDENT RATES BASED ON 1961–62 ACCIDENT EXPERIENCE FOR MALES[a]

Accidents 1961–62	1963 accident rates	
	Empirical	Estimated
0	.0721	.0723
1	.1112	.1105
2+	.1454	.1529

[a] *Source*: [3].

4. MULTIPLE POISSON REGRESSION ANALYSIS

For the sake of simplification in the development that follows, let $t = 1$ in Equation (1.1). Then, from (1.1) and (3.1), the probability that the jth individual in a sample will be involved in n_j accidents during the next unit of time is given by

$$p(n_j) = \frac{\exp[-\sum_{i=0}^{k} \beta_i x_{ij}](\sum_{i=0}^{k} \beta_i x_{ij})^{n_j}}{n_j!}, \tag{4.1}$$

$$n_j = 0, 1, 2, \cdots, \qquad \sum_{i=0}^{k} \beta_i x_{ij} > 0.$$

With respect to a sample of size s, the likelihood function is

$$L = \prod_{j=1}^{s} \frac{\exp[-\sum_{i=0}^{k} \beta_i x_{ij}](\sum_{i=0}^{k} \beta_i x_{ij})^{n_j}}{n_j!}$$

Taking the natural logarithm we obtain

$$\ln L = -\sum_{j=1}^{s}\sum_{i=0}^{k} \beta_i x_{ij} + \sum_{j=1}^{s} n_j \ln(\sum_{i=1}^{k} \beta_i x_{ij}) - \sum_{j=1}^{s} \ln n_j!$$

Differentiating with respect to β_i, $i = 0, 1, 2, \cdots, k$, we get

$$\frac{\partial \ln L}{\partial \beta_i} = -\sum_{j=1}^{s} x_{ij} + \sum_{j=1}^{s} \frac{x_{ij} n_j}{(\sum_{i=0}^{k} \beta_i x_{ij})}$$

On setting the $k+1$ partials equal to zero, the system of maximum likelihood normal equations obtained is

$$\sum_{j=1}^{s} \frac{x_{ij} n_j}{(\sum_{i=0}^{k} \hat{\beta}_i x_{ij})} = \sum_{j=1}^{s} x_{ij}, \quad i = 0, 1, \cdots, k. \tag{4.2}$$

Jorgenson [7] showed that a solution to the set of Equations (4.2) can be obtained by using an iterative weighted least-squares procedure. If N_j is the random variable having distribution (4.1), then the parameter λ associated with the jth individual is

$$\lambda_j = E(N_j) = \sum_{i=0}^{k} \beta_i x_{ij} = \mathrm{Var}(N_j).$$

In matrix notation,

$$\lambda = E(N) = X\beta \quad \text{and} \quad \mathrm{Cov}(N) = V$$

where λ, N and β are $s\times 1$, $s\times 1$ and $(k+1)\times 1$ vectors. X is an $s\times(k+1)$ matrix having the values of the criterion variables as elements and V is an $s\times s$ diagonal matrix with elements $v_j=\sum_{i=0}^{k}\beta_i x_{ij}$. It is well known (e.g., see [5]) that the minimum variance linear unbiased estimator for β is

$$\hat{\beta}=(X'V^{-1}X)^{-1}X'V^{-1}N.$$

Also, according to general linear model theory, if x_j is the vector of criterion variables corresponding to the jth individual,

$$E(x_j'\hat{\beta})=\lambda_j \quad \text{and} \quad \operatorname{Var}(x_j'\hat{\beta})=x_j'(X'V^{-1}X)^{-1}x_j.$$

Unfortunately, since β is unknown, the matrix V is unknown. Our problem then is to obtain an estimate of V which in turn gives us an estimate of β. Following Jorgenson [7], we let $\hat{V}_m$ denote the estimate of V obtained on the mth iteration and we let the corresponding estimate of β be

$$b_m=(X'\hat{V}_m^{-1}X)^{-1}X'\hat{V}_m^{-1}n \tag{4.3}$$

where n is the vector of values corresponding to the random vector N. Let $\hat{V}_0$ be the $s\times s$ identity matrix and define

$$\hat{V}_{m+1}=\operatorname{diag}[x_1'b_m,\ x_2'b_m,\ \cdots,\ x_s'b_m].$$

The iterations are continued until convergence is realized, i.e., $b_{m+1}=b_m$. Denote this equality vector by b. Then

$$b=(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}n \tag{4.4}$$

where $\hat{V}$ is the equality matrix $\hat{V}_{m+1}=\hat{V}_m$. As our final estimate of λ_j we use

$$\hat{\lambda}_j=x_j'b \tag{4.5}$$

and as an estimate of the variance of $\hat{\lambda}_j$ we may use

$$\operatorname{Var}(\hat{\lambda}_j)=x_j'(X'\hat{V}^{-1}X)^{-1}x_j. \tag{4.6}$$

As a result of using $\hat{V}$ in place of the unknown matrix V, the estimate (4.5) for λ_j is not unbiased and the variance of the corresponding estimator is unknown. However Jorgenson [7] points out that (4.4) is best asymptotically normal (BAN) and that the iterative procedure converges provided that $\hat{V}_m$ and $(X'\hat{V}_m^{-1}X)^{-1}$ are positive definite for all m.

Work by Wald [8] provides a theoretical basis for testing

$$H_0: L\beta=\gamma \tag{4.7}$$

where L is a known $\ell\times(k+1)$ matrix of rank $\ell\leq k+1$ and γ is a specified vector of constants. The appropriate test statistic

$$W=(Lb-\gamma)'[L(X'\hat{V}^{-1}X)^{-1}L']^{-1}(Lb-\gamma) \tag{4.8}$$

is asymptotically distributed as chi square with ℓ degrees of freedom. This can be used to test such hypotheses as

$$H_0:\beta_i=0 \quad \text{and} \quad H_0:\lambda=x'\beta=\lambda_0.$$

Using formula (4.3) to compute the vector b_{m+1} is not feasible when the sample size s is large, as in the example of the next section where $s=86{,}726$. In this study [9], b_{m+1} was calculated for $m=0, 1, 2, \cdots$, by applying a weight of

$$\left(\sum_{i=0}^{k} b_{i(m)}x_{ij}\right)^{-1/2}$$

to the data and then using a standard least-squares regression program. Here $b_{i(m)}$ is the ith element in the vector b_m. The reader is reminded that this procedure is simply the method of weighted least squares where the weights are the reciprocals of the standard deviations of the dependent variables and the system of normal equations is

$$\sum_{j=1}^{s}\frac{x_{ij}\left(n_j-\sum_{i=0}^{k}b_{i(m+1)}x_{ij}\right)}{\sum_{i=0}^{k}b_{i(m)}x_{ij}}=0,\quad i=0, 1, \cdots, k.$$

It is readily seen that this system reduces to system (4.2) when $b_{i(m+1)}=b_{i(m)}$ for all $i=0, 1, 2, \cdots, k$.

5. NUMERICAL EXAMPLE

In this section we summarize the results of the multiple Poisson regression technique discussed in the previous section as applied to the sample of nearly 87,000 male drivers in the California study. With n representing the number of accident involvements during 1963, the selected criterion variables were:

- x_1 = the natural logarithm of the traffic density index of the county in which the driver resides,
- $x_2=5/(\text{age}-13)$,
- x_3 = the number of countable convictions incurred during years 1961–62,
- x_4 = the number of accident involvements incurred during years 1961–62,
- x_5 = the number of noncountable convictions incurred during years 1961–62.

Convergence to seven decimal places in the b vector was achieved on the fifth iteration. The final estimation function for λ and the estimate of the covariance matrix of the β estimator are:

$$\lambda=0.00274+0.00909x_1+0.0532x_2+0.0223x_3+0.0216x_4+0.0169x_5$$

$$(X'\hat{V}^{-1}X)^{-1}=10^{-4}\begin{bmatrix} 0.1981 & -0.0384 & -0.0754 & 0.0041 & 0.0021 & -0.0024\\ -0.0384 & 0.0085 & -0.0004 & -0.0012 & -0.0013 & 0.0007\\ -0.0754 & -0.0004 & 0.4058 & -0.0137 & -0.0057 & -0.0162\\ 0.0041 & -0.0012 & -0.0137 & 0.0142 & -0.0052 & -0.0050\\ 0.0021 & -0.0013 & -0.0057 & -0.0052 & 0.0654 & -0.0030\\ -0.0024 & 0.0007 & -0.0162 & -0.0050 & -0.0030 & 0.0623 \end{bmatrix}$$

It is instructive to see how the final estimate for β_i, namely, b_5 compares with the initial, or unweighted least-squares estimate, b_0, and the first weighted least-squares estimate, b_1. This is shown in Table 4.

Table 4. COMPARISON OF FINAL ESTIMATE FOR β WITH FIRST TWO ESTIMATES

Variable	b_0	b_1	b_5
x_0	-.0020254	.0028429	.0027445
x_1	.0102197	.0090929	.0090912
x_2	.0566802	.0530796	.0532423
x_3	.0197855	.0222332	.0222984
x_4	.0221827	.0214463	.0215772
x_5	.0177242	.0168427	.0168559

Table 5. ESTIMATES OF λ AND ITS STANDARD DEVIATION FOR INDIVIDUAL MALE DRIVERS

Traffic density	Age	Convictions	Accidents	Non-countable convictions	$\hat{\lambda}$	$\sqrt{\text{Var}(\hat{\lambda})}$
10	60	0	0	0	.0893	.0023
50	60	1	0	1	.0831	.0027
150	60	1	1	1	.1147	.0034
10	40	2	0	0	.0781	.0032
50	40	0	1	0	.0697	.0027
150	40	0	0	0	.0581	.0011
10	20	1	1	0	.1056	.0045
50	20	0	0	0	.0763	.0035
150	20	3	2	1	.2111	.0099

In Table 5, the values of $\hat{\lambda}$ and its estimated standard deviation are given for selected values of the criterion variables.

Finally, we illustrate the use of the Wald asymptotic test statistic (4.8). Suppose we wish to test the hypothesis

$$H_o: \beta_4 = 0.$$

Referring to statement (4.7), here

$$L = [0, 0, 0, 0, 1, 0] \quad \text{and} \quad \gamma = 0.$$

Hence

$$W = ((2.16 \times 10^{-4})^2/0.0654 \times 10^{-4}) = 71.35.$$

Comparing with 6.64, the .01 point of the $\chi^2(1)$ distribution, we conclude that $\beta_4 \neq 0$.

As a second example, suppose it is of interest to test the hypothesis

$$H_o: \lambda = 0.2000$$

for those drivers belonging to the class represented by the last line in Table 5. In this instance,

$$L = [1, \ln 150, 5/7, 3, 2, 1] \quad \text{and} \quad \gamma = 0.2000.$$

Calculation of the W test statistic gives us

$$W = ((.0131)^2/(.0059)^2) = 4.93$$

which falls between the .05 and the .025 points of the $\chi^2(1)$ distribution.

REFERENCES

[1] Arbous, A. G., and J. E. Kerrich, "Accident Statistics and the Concept of Accident-Proneness," *Biometrics*, 7 (December 1951), 340–432.

[2] Bates, G. E., and J. Neyman, "Contributions to the Theory of Accident-Proneness," *University of California Publications in Statistics*, I (April 1952), 215–75.

[3] California Department of Motor Vehicles, *The California Driver Record Study*, Parts 1–9, Sacramento, 1964–67.

[4] Edwards, C. B., and J. Gurland, "A Class of Distributions Applicable to Accidents," *Journal of the American Statistical Association*, 56 (September 1961), 503–17.

[5] Goldberger, A. S., *Economic Theory*, New York: John Wiley & Sons, Inc., 1964.

[6] Greenwood, M. and G. U. Yule, "An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents," *Journal of the Royal Statistical Society*, 83 (March 1920), 255–79.

[7] Jorgenson, Dale W., "Multiple Regression Analysis of a Poisson Process," *Journal of the American Statistical Association*, 56 (June 1961), 235–45.

[8] Wald, A., "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," *Transactions American Mathematical Society*, 54 (November 1943), 426–82.

[9] Weber, D. C., "A Stochastic Model for Automobile Accident Experience," Institute of Statistics Mimeograph Series No. 651, North Carolina State University at Raleigh, 1970.